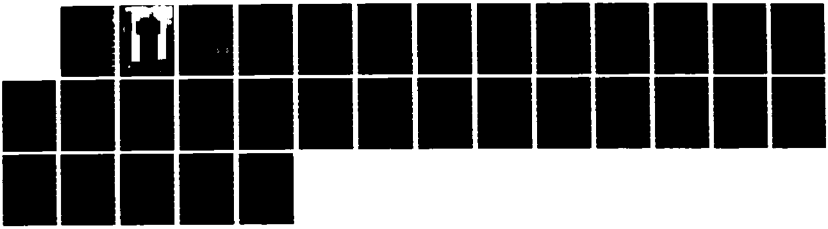
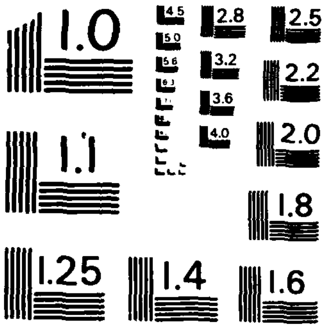
AD-A169 828 THE BASKET METHOD AND ITS ROLE IN DESIGN AND MODEL ORIENTED APPROACHES TO..(U) CLEMSON UNIV SC DEPT OF MATHEMATICAL SCIENCES R J DOMANGUE ET AL. JAN 83 1/1

UNCLASSIFIED TR-408 N00014-75-C-0451 F/G 12/1 NL





MICROCOPY RESOLUTION TEST CHART

NATIONAL BUREAU OF STANDARDS - 1963 - A

THE BASKET METHOD AND ITS ROLE IN DESIGN AND MODEL ORIENTED APPROACHES TO SURVEY SAMPLING

R. J. DOMANGUE
AND
K. T. WALLENIUS

DEPARTMENT
OF
MATHEMATICAL
SCIENCES
CLEMSON UNIVERSITY
Clemson, South Carolina

THE BASKET METHOD AND ITS ROLE IN DESIGN AND MODEL ORIENTED APPROACHES TO SURVEY SAMPLING

R. J. DOMANGUE

AND

K. T. WALLENIUS

DEPARTMENT OF MATHEMATICAL SCIENCES

CLEMSON UNIVERSITY

TECHNICAL REPORT #408

JANUARY, 1983

N144

THIS WORK WAS SUPPORTED IN PART BY THE OFFICE OF NAVAL RESEARCH UNDER CONTRACT N00014-75-C-0451.

Table of Contents

I. Introduction and Summary

A controversy currently exists among statisticians involving the foundations for inference based on a sample from a finite population. This controversy centers on the role of the sampling design and involves fundamental questions in statistical theory. There are two main viewpoints. In the classical or traditional approach, the sampling design plays a major role in inference since this approach is based on the Randomization Principle. Attention usually focuses on statistical properties of estimators (e.g. bias, mean square error) relative to a given design. In the model-based approach, inference is made within a model framework, a relatively new idea in survey sampling yet a very common one in other statistical procedures. In this approach the main concern is optimality of estimators with respect to the given model, the sampling design playing a secondary role.

In this paper we describe the two approaches and discuss their advantages and disadvantages. Secondly, a review of some of the work of Richard Royall is given, Royall being one of the staunchest supporters of the model-based approach. Lastly we describe a sampling procedure called the "basket method" and examine its relationship to the two approaches.

II. The Problem and the Two Approaches

A. The Problem

The basic problem in survey sampling is the following: A population of interest contains N units, labelled 1, ..., N. Associated with each unit is an unknown value y_i. The vector $y = (y_1, ..., y_N)$ may be considered a parameter of the population with interest usually focusing on some function of y such as the population total

$$T = \Sigma_{i=1}^{N} y_i$$

We will consider problems in which an auxiliary value x_i, usually a size value, is known for each unit. Presumably the x's contain some information about the y's. That informational relationship is characterized and exploited for inference in the model-based approach but ignored in the classical approach. A sample from the population is selected, the y's are observed and used to estimate the parameter of interest with the x's possibly providing aid in selection and/or estimation.

B. Classical Approach

1. Description

In the classical approach, the y's are treated as fixed but unknown constants. A sample s of size n is selected according to some sampling design P. In mathematical terms, P is a probability function on the collection S of all subsets of size n which can be formed from the indices {1, 2, ..., N} with P(s) denoting the probability of selecting those units whose labels are in s.

The data $d = \{(i,y_i), i \varepsilon s\}$ is observed and T is estimated with some estimator $\hat{T}$. Note that the labels are considered as part of the data since many estimators in survey sampling depend on the labels. Properties such as bias and MSE (mean squared error) are based on the probability distribution generated by the sampling design P, viz.

$$\text{bias} = E_P[\hat{T} - T] = \sum_S P(s)(\hat{T} - T)$$

$$\text{MSE} = E_P\ [(\hat{T} - T)^2] = \sum_S P(s)(\hat{T} - T)^2$$

2. Estimators and Designs

Let

$$\bar{y}_s = \sum_s y_i/n \text{ and } \bar{x}_s = \sum_s x_i/n$$

denote the sample mean of the y's and x's, respectively. Let

$$X = \Sigma_{i=1}^{N} x_i$$

be the population total of the x's and

$$\bar{X} = \Sigma_{i=1}^{N} x_i/N$$

the population mean. Finally let Π_i be the probability that unit i is selected in the sample for any arbitrary design P.

Some common sampling designs and estimators are given below:

Designs:

a. simple random sampling (SRS):

$$\Pi_i = \frac{n}{N} \text{ for each } i = 1, \ldots, N$$

b. probability proportional to size (PPSX):

$$\Pi_i = \frac{nx_i}{N\bar{X}} \text{ (available only if } nx_i < N\bar{X} \text{ for all i)}$$

c. probability proportional to aggregate size (PPAS):

$$\text{PPAS}(s) = \sum_s x_i/N'X \qquad N' = \binom{N-1}{n-1}$$

Estimators:

a. $\hat{T}_E = N\bar{y}_s$ (expansion estimator)

$$\hat{T}_R = \Sigma_{k=1}^{N} (\sum_s y_i/\sum_s x_i)\, x_k = (\bar{y}_s/\bar{x}_s) * \Sigma_{k=1}^{N} x_k \qquad \text{(ratio estimator)}$$

$$\hat{T}_{HT} = \Sigma_{k=1}^{N} (\frac{1}{n} \sum_s (y_i/x_i))\, x_k \qquad \text{(Horvitz-Thompson estimator)}$$

A sampling design P combined with an estimator $\hat{T}$ is called a sampling strategy and denoted by the pair $(P,\hat{T})$. Consideration is usually given to strategies $(P,\hat{T})$ in which $\hat{T}$ is unbiased with respect to the design P (called a P-unbiased strategy). It is well known that ($\hat{T}_E$, SRS), ($\hat{T}_{HT}$, PPSX) and ($\hat{T}_R$, PPAS) are P-unbiased strategies.

3. Arguments For and Against

"Model robustness" in statistics refers to a procedure whose performance does not seriously deteriorate under certain types of departures from the assumed model. Proponents of the classical approach argue that design based inference is robust in the sense that no probabilistic assumptions need be made. The estimator $\hat{T}_E$ is unbiased under SRS no matter how the y's are distributed. They also argue that the use of randomization in selecting samples averages out effects of unobserved or unknown random variables and safeguards against selection bias.

However, some statisticians point to "negative aspects" of the classical model, for instance, the non-existence of unbiased minimum variance estimation [2]. Also the likelihood function arising from this model is informative only in a trivial sense [1]. Given the data $d = \{(i, y_i), i \varepsilon s\}$, all N-vectors $y^* = (y_1^*, \ldots, y_N^*)$ such that $y_i^* = y_i$, $i \varepsilon s$, have the same likelihood, namely P(s). No unique maximum likelihood estimator exists. When the likelihood principle is applied to survey sampling it implies that the sampled y's should give the same inference no matter what the design. Also in the classical approach, s is an ancillary statistic and thus any inference consistent with the conditionality principle should be conditioned on s. But the conditional distribution of the data given s is degenerate. These aspects prompted the use of probability models as a basis for inference in survey sampling.

C. Model-Based or Superpopulation Approach

1. Description

In the model-based approach, the numbers $y_1, \ldots, y_N$ are treated as realizations of random variables $Y_1, \ldots, Y_N$ characterized by some model ξ. Hence the population at hand can itself be considered a sample from a "superpopulation". In this approach the design is relegated to a secondary role and emphasis is placed on estimators that are "good" with respect to the model no matter how s was selected.

Once a sample s is selected and the y's of the sampled units observed, the quantity T to be estimated can be written as

$T = \sum_s y_i + \sum_{\tilde{s}} y_i$, the sum of the known

y's plus the sum of the values of the random variables for those units not in the sample. Thus estimating T becomes a problem of <u>predicting</u> $\sum_{\tilde{s}} Y_i$. The sampled units provide information on the model parameters in ξ which can be used to predict $\sum_{\tilde{s}} Y_i$. The predictor usually takes the form

$$\hat{T} = \sum_s y_i + U,$$

where U is the predictor for $\sum_{\tilde{s}} Y_i$. Thus inference is based on the model ξ, contrary to the Randomization Principle. Given a sample s, bias and mean square error (MSE) are defined as

$$\text{bias} = E_\xi[\hat{T} - T]$$

$$\text{MSE} = E_\xi[(\hat{T} - T)^2]$$

Under the model-based approach, a good predictor $\hat{T}$ is one which is optimal in some sense under the model no matter which sample is selected. A standard optimality criterion, minimum mean square error, calls for selecting a predictor U which minimizes MSE. This criterion has received exhaustive attention in the statistics literature and solutions exist for a wide collection of models. For example, suppose $x_i' = (x_{i0}, x_{i1}, \ldots, x_{ir})$ with $x_{i0} \equiv 1$ is a vector of known auxiliary information and ξ is the linear regression model in which the Y's are independent with

$$E_\xi[Y_i] = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_r x_{ir} = x_i'\beta = \mu_i,$$

$$\mathrm{Var}_\xi[Y_i] = \sigma^2 v_i$$

Here $\beta' = (\beta_0, \beta_1, \ldots, \beta_r)$ and σ^2 are unknown parameters and v_i is known. In this case linear least squares prediction can be applied. If we restrict $\hat{T}$ to be linear in the sample y's and ξ-unbiased, i.e. $E_\xi[\hat{T}] = E_\xi[T]$ then $E_\xi[(\hat{T}-T)^2]$ is minimized by taking

$$\hat{T} = \hat{T}_{\xi\text{-BLU}} = \sum_s y_i + U^*$$

where U* is the BLU (best linear unbiased) estimator of $E_\xi[\sum_{\tilde{s}} Y_i]$ obtained from generalized least squares. Thus, for any particular sample s, $\hat{T}_{\xi\text{-BLU}}$ minimizes mean square error, but the value of this minimum will, in general, depend on which sample s is selected so that the sample design may be important even in the modal based approach. This suggests tailoring the sample design to the estimator (in contrast to the classical approach of tailoring the estimator to the design) and often leads to a degenerate sampling design which places all its mass on the "best" sample. Adherents to the classical approach argue against such a "purposive" design since randomization is eliminated as well as estimates of variance which result from randomization.

2. Results of Royall et. al.

Richard Royall has been one of the main advocates of the prediction approach to survey sampling. Some of his work is reviewed below using the notation in [5].

Let $\xi(\delta_0, \delta_1, \ldots, \delta_J : v(x))$ refer to the polynomial regression model wherein $Y_1, \ldots, Y_N$ are independent random variables with

$$Y_i = \delta_0\beta_0 + \delta_1\beta_1 x_i + \ldots + \delta_J\beta_J x_i^J + \varepsilon_i[v(x_i)]^{1/2}$$

The δ's are indicator variables allowing for the inclusion (δ_j=1) or deletion (δ_j=0) of the term $\beta_j x^j$ for j = 1, 2, ..., J. $\varepsilon_1, \ldots, \varepsilon_N$ are independent random variables each having mean 0 and variance σ^2; $v(\bullet)$ is a known function. Hence under ξ

$$E_\xi[Y_i] = \delta_0\beta_0 + \delta_1\beta_1 x_i + \ldots + \delta_J\beta_J x_i^J$$

and

$$\text{Var}_\xi[Y_i] = \sigma^2 v(x_i)$$

As an example, consider the model $\xi_1(0,1:x)$. Here $Y_1, \ldots, Y_N$ are independent random variables with

$$E_{\xi_1}[Y_i] = \beta x_i$$

$$\text{Var}_{\xi_1}[Y_i] = \sigma^2 x_i$$

β and σ^2 unknown.

a. Results under ξ_1 [6]

1. For any given sample s

$$\hat{T}_{\xi_1\text{-BLU}} = \hat{T}_R = \Sigma_{k=1}^{N} (\Sigma_s y_i / \Sigma_s x_i)\, x_k \qquad \text{(ratio estimator)}$$

and for any given design P,

$$E_{\xi_1} E_P [(\hat{T}_R - T)^2] = E_P[(N/n)(N-n)\sigma^2 \bar{X}(\bar{x}_s/\bar{x}_{\bar{s}})]$$

2. The optimal strategy is $(\hat{T}_R, P^*)$ where

$$P^*(s) = \begin{cases} 1 & \text{if } s = s^* \\ 0 & \text{if } s \neq s^* \end{cases}$$

where s* denotes the set of labels with the largest x-values.

3. $\hat{T}_R$ can be extremely biased if s* is used and ξ_1 is not the true model. For example if $\xi_2(1,1:x)$ is the true model and $s = s^*$,

$$E_{\xi_2}[\hat{T}_R - T] = N\beta_0(\bar{X} - \bar{x}_{s^*})/\bar{x}_{s^*} \neq 0$$

Royal defines a "balanced sample" s as one whose moments match those of the population. Specifically, s is said to be a "balanced sample" of order J, denoted s(J), if

$$\bar{x}_s^{(j)} = \bar{X}^{(j)} \qquad j = 1, \ldots, J$$

where

$$\bar{x}_s^{(j)} = \sum_s x_i^j/n, \quad \bar{X}^{(j)} = \Sigma_{i=1}^N x_i^j/N$$

4. If s = s(J), $\hat{T}_R$ is unbiased under $\xi(\delta_0, \delta_1, \ldots, \delta_J; v(x))$.

5. When a balanced sample s(J) is selected, then $\hat{T}_{\xi\text{-BLU}} = \hat{T}_E = N\bar{y}_s$, the expansion estimator, under the model $\xi(\delta_0, \delta_1, \ldots, \delta_J: v(x))$. Here

$$v(x) = \Sigma_{j=0}^J \delta_j a_j x^j$$

and $a_0, a_1, \ldots, a_J$ are any positive constants.

Therefore, Royall suggests the use of balanced samples to achieve robustness in model-based inference, giving up a small amount of efficiency for protection against model departure. He advises, however, the use of some sort of randomization in selecting balanced samples as a means of balancing on variables which are not explicitly considered in the model. With today's high speed computers one could randomly choose from (approximately) balanced samples by selecting simple random samples until one is obtained that meets a pre-assigned criteria. Such an approach is used by Royall and Cumberland in [4].

3. Robust Variance Estimation

Royall strongly believes that inference should be model-based and conditioned on s. Departures from assumed models become a concern in this approach so Royall concentrated his work on robust model-based procedures such as balanced sampling. Royall and Eberhardt [5] have also dealt with robust variance estimation procedures.

Now for the advocate of model-based inference, the estimate of the precision of an estimator is the error variance

$$\mathrm{Var}_{\xi}\,(\hat{T} - T) = E_{\xi}\,[(\hat{T} - T)^2]$$

The error variance usually involves unknown parameters and thus must be estimated. This estimation relies heavily on the variance structure of the adopted model, which can be biased if the true model has a different structure. Thus what is needed is a robust procedure for estimating this error variance. We take the model ξ_1 as an example.

As previously mentioned, under model $\xi=\xi_1\ (0,1:x)$, for a given sample s,

$$\hat{T}_R = \Sigma_{k=1}^{N}\ (\Sigma_s\, y_i/\Sigma_s\, x_i)\ x_k$$

is ξ_1-BLU and the error variance is

$$\mathrm{Var}_{\xi_1}(\hat{T}_R-T) = E_{\xi_1}[(\hat{T}_R-T)^2] = \sigma^2(\Sigma_{\tilde{s}}x_i/\Sigma_s x_i)\ \Sigma_{i=1}^{N}\ x_i$$

An estimator V_L of $\mathrm{Var}_\xi(\hat{T}_R - T)$ can be obtained from weighted least squares and is

$$V_L = (N/n)(N-n)[(\sum_s d_i^2/x_i)/(n-1)]\ \bar{x}_{\bar{s}}\bar{X}/\bar{x}_s$$

where

$$d_i = y_i - (\bar{y}_s/\bar{x}_s)x_i$$

V_L is ξ_1-unbiased for all samples s but only if $\mathrm{Var}[Y_i]\ \alpha\ x_i$. V_L can be badly biased if the model fails in that respect.

The usual estimator of variance of $\hat{T}_R$ when simple random sampling is used is

$$V_C = (N/n)(N-n)\ \sum_s d_i^2/(n-1)$$

Royall and Eberhardt [5] derived two variance estimators V_1 and V_2 which are model unbiased for all samples when variance is proportional to size, but which is asymptotically unbiased for quite general variance structures:

$$V_1 = V_C\ (\bar{x}_{\bar{s}}\bar{X}/\bar{x}_s^2)(1 - v_s^2/n)^{-1}$$

where

$$v_s^2 = (n-1)^{-1}\ \sum_s (x_i - \bar{x}_s)^2/\bar{x}_s^2$$

$$V_2 = (N/n^2)(N-n)(\bar{x}_{\bar{s}}-\bar{X}/\bar{x}_s^2)\sum_s d_i^2/[1-(x_i/n\bar{x}_s)]$$

In an empirical study of six populations in which the model ξ_1 might apply, Royall and Cumberland [4] compare the above estimators. Their general conclusions are that V_1 and V_2 are superior to V_C in simple random samples, and even these were reliable only in well-balanced samples. The estimator V_C depends on the value of $\bar{x}_s$ and is greatly biased if $\bar{x}_s$ is very different from $\bar{X}$. V_1, V_2, and V_C are approximately equal when $\bar{x}_s = \bar{X}$ and inference is much better under these samples. The estimator V_L, although unbiased under ξ_1 for every s, was not robust enough to perform well in the studied populations.

III. The "Basket Method" for Selecting Balanced Samples.

A. Background

About the time that Royall and Herson were developing the concepts of robust estimation within the framework of a model-based approach [6], Wallenius was investigating an application of sampling in the area of price estimation [8]. The scenario for this application is described in detail in [10]. Briefly, the units comprising the population are price proposals from a sole-source contractor with whom the government is doing business. Associated with the ith unit is the contractor's proposed price x_i and an unknown price y_i which will be determined through a time consuming processes of government price analysis and negotiation with the contractor. If we assume the contractor has a fairly good idea of what the latter price should be, he could set the proposed price in accordance with any "padding strategy" which suited his purpose. Thus we are dealing with a situation involving competition with large sums of money at stake in which one of the game players, the contractor, can control the state of nature. It is up to the other player, the government, to carefully analyze each proposal and fully prepare for the negotiation phase in order to avoid overpayment. The process can be quite involved both technically and time-wise. It is fairly common to find a large backlog of proposals awaiting analysis and negotiation. Since processing this backlog in an expeditious manner is in the best interest of both the contractor and the government, the situation seemed appropriate for a statistical treatment. The previously mentioned element of competition required special consideration: most classical procedures were vulnerable to gamesmanship and, by the same token, it would be foolish to base inference on any structured model for the relationship between proposed and negotiated prices. In the next section we describe a statistical tool called the "basket method" developed to handle this situation. It combines the concepts of randomization and balanced sampling by partitioning the population into a collection of balanced subsets from which one is selected at random.

B. Description

Suppose that the population under study is to be partitioned into K disjoint subsets $b_1, b_2, \ldots, b_K$, not necessarily of equal size. One subset is selected at random, the y's observed, and T estimated on the basis of that sample. If it were possible to accomplish this partitioning in such a way that

$$\sum_{b_j} y_i = \sum_{b_k} y_i = (1/K) \sum_{i=1}^{N} y_i$$

for every pair $j \neq k$, then the estimator $\hat{T}_B = K^* \sum_b y_i$ (b being the sample selected at random) would equal T no matter which sample was selected. But the y's are unknown in advance so this type of balance cannot be expected. If we assume that the y's are characterized by some probability model ξ, then a reasonable approximation to the above procedure would be to partition the units into subsets such that

$$\sum_{b_j} \hat{Y}_i = \sum_{b_k} \hat{Y}_i = (1/K) \sum_{i=1}^{N} \hat{Y}_i$$

for every pair $j \neq k$ where $\hat{Y}_i = E_\xi[Y_i]$ is the predicted Y_i under the model ξ. Then choose a sample and estimate T with $\hat{T}_B = K \sum_b Y_i$. If we let P^B denote the design that selects one sample at random from the K samples, and assume the Y's are independent under ξ, then it can be shown that, with the strategy $(P^B, \hat{T}_B)$, the optimal allotment of units to samples under ξ is one in which

$$\sum_{b_j} \hat{Y}_i = \sum_{b_k} \hat{Y}_i$$

for every $j \neq k$. Specifically $E_\xi E_{P^B}[(\hat{T}_B - T)^2]$ is minimized for such an allotment which we shall call balancing on total predicted Y in the sample. As in the previous case, however, balancing on total predicted Y cannot be carried out in practice (since $\hat{Y}_i$ may depend on unknown parameters in ξ). Fortunately, for a wide class of practical models ξ, balancing on total predicted Y can be achieved even though the model parameters are unknown.

Consider the model $\xi_1(0,1: x)$ previously mentioned. Under this model $E_{\xi_1}[Y_i] = \beta x_i$ where β is unknown and x_i known. The "basket method" can achieve the balance on total predicted Y in the sample under this model.

The "basket method" [10] is a procedure whereby the population of N units is partitioned into K samples called "baskets" of size $n = [N/K]$ or $n = [N/K] + 1$ (where $[\bullet]$ denotes the greatest integer function) such that

$$\sum_{bsk(e)} x_i \simeq \sum_{bsk(f)} x_i \simeq \frac{1}{K} \sum_{i=1}^N x_i$$

for any two baskets e and f. A basket is selected at random and T is estimated by

$$\hat{T}_R = (\sum_{bsk(e)} y_i / \sum_{bsk(e)} x_i) \sum_{i=1}^N x_i \simeq K \sum_{bsk(e)} y_i$$

Note that

$$\sum_{bsk(e)} E_{\xi_1}[Y_i] = \beta \sum_{bsk(e)} x_i \approx \beta \sum_{bsk(f)} x_i = \sum_{bsk(f)} E_{\xi_1}[Y_i]$$

for any two baskets e and f and the desired balance on total predicted Y in the sample is achieved, whatever be the value of β.

Briefly the basket formation algorithm works as follows. If a sample size of approximately n is desired then $K = [N/n]$ baskets are formed. The units are arranged in decreasing order on x and labelled 1, ..., N with the unit having the largest x value labelled 1, 2nd largest labelled 2 and so on. Starting with the first K units (K largest x's) place one unit in each basket. The remaining units are partitioned into successive groups of K units each. A group of K units is assigned to the K baskets by the following rule: Compute the basket totals and arrange the baskets in order of increasing totals [10]. Then assign the K units, one-per-basket, in sequential fashion with the largest unassigned unit being placed in the basket with the smallest total. If N/K is not an integer then the last group (smallest x's) will not contain K units, the result being that some baskets will have one fewer number of units. As mentioned earlier, actual basket size is either $[N/K]$ or $[N/K] + 1$. The initial basket formation should result in nearly equal basket totals on x, but a swapping algorithm is used to bring basket totals into even closer agreement. Experience gained by applying this simple algorithm to real populations indicates the basket method results in nearly identical basket totals. It is possible, of course, to construct a population for which the basket method does not yield a good balance on total x but we have not encountered such a situation in practice.

Empirical evidence shows that the basket formation algorithm, while designed to achieve good balance on $\sum_{bsk(e)} x_i$, does surprisingly well in balancing higher moments as well. This is due to the nature of the method. Hence we may expect to achieve approximate balancing of total predicted Y in the sample under higher degree polynomial models. For example, suppose that $E[Y_i] = \beta_0 + \beta_1 x_i + \beta_2 x_i^2$. If

$$\sum_{bsk(e)} x_i \approx \sum_{bsk(f)} x_i \text{ and } \sum_{bsk(e)} x_i^2 \approx \sum_{bsk(f)} x_i^2$$

for any pair of baskets e and f then

$$\sum_{bsk(e)} E[Y_i] \approx \sum_{bsk(f)} E[Y_i].$$

Since the sample size n may differ by one unit from basket to basket, inclusion of an intercept term β_0 can have a small adverse effect on the total predicted y balance. Supposing basket f contains n units and basket e contains (n+1) units,

$$\sum_{bsk(e)} E(Y_i) = (n+1)\beta_0 + \beta_1 \sum_{bsk(f)} x_i \; \beta_2 \sum_{bsk(f)} x_i^2 \approx \beta_0 + \sum_{bsk(f)} E(Y_i).$$

In practice, β_0 is usually <u>very</u> small relative to $\sum_{bsk(f)} E(Y_i)$ so the problem is inconsequential.

Now suppose we have adopted a polynomial model $\xi(\delta_0, \delta_1, \ldots, \delta_J: v(x))$, i.e.

$$Y_i = p(x_i) = \delta_0\beta_0 + \delta_1\beta_1 x_i + \ldots + \delta_J\beta_J x_i^J + \varepsilon_i[v(x_i)]^{1/2}$$

but the true model is

$$Y_i = f(x_i) + \varepsilon_i[w(x_i)]^{1/2}$$

where f and w are arbitrary unknown functions. Assuming f can be adequately approximated with a Taylor series polynomial p* through the Jth term, we have

$$E[Y_i] = f(x_i) \approx p^*(x_i) = \beta_0^* + \beta_1^* x_i + \ldots + \beta_J^* x_i^J$$

Even though the coefficients β_j^* are unknown, the basket method algorithm will result, approximately, in

$$\sum_{bsk(e)} p^*(x_i) \approx \sum_{bsk(f)} p^*(x_i)$$

so that, again,

$$\sum_{bsk(e)} E[Y_i] \approx \sum_{bsk(f)} E[Y_i]$$

for any two baskets e and f.

C. Relation to Royall's Work

Now since

$$\sum_{bsk(e)} x_i \approx \frac{1}{K} \Sigma_{k=1}^{N} x_k$$

for every basket e we have

$$\bar{x}_{bsk(e)} = \sum_{bsk(e)} x_i/n \approx \frac{1}{nK} \Sigma_{k=1}^{N} x_k = (N/nK)\bar{X} = \begin{cases} \bar{X} \text{ if } (N/K) = [N/K] \\ (1\pm\delta)\bar{X} \text{ if } (N/K) \neq [N/K] \end{cases}$$

where $0 \leq \delta \leq \frac{1}{n}$. Thus $\bar{x}_{bsk(e)} \approx \bar{X}$ and we have a Royall "balanced sample" of degree 1. Similarly, if

$$\sum_{bsk(e)} x_i^j \approx \sum_{bsk(f)} x_i^j \approx \frac{1}{K} \Sigma_{i=1}^{N} x_i^j$$

for $j = 2, \ldots, J$, then

$$\bar{x}^{(j)}_{bsk(e)} \approx \bar{x}^{(j)}_{bsk(f)} \approx \bar{X}^{(j)}$$

and the sample is a "balanced sample of degree J." As we have seen previously, estimators which are optimal relative to certain model assumptions have some nice robustness properties if we take the trouble of finding a balanced sample. The ratio estimator is unbiased under a number of polynomial models when the sample is balanced and also optimal is some cases.

The basket method is also related to some of Royall's work in balancing stratified samples. Let the population of N units be stratified into H strata as follows: the N_1 units whose x-values are smallest form stratum 1, the next N_2 smallest units form stratum 2, and so on. For h = 1, 2, ..., H, a sample s_h of size n_h is selected from strata h and the stratum totals are estimated by

$$\hat{T}_h = (\Sigma_{s_h} y_{hi}/\Sigma_{s_h} x_{hi})\Sigma_{k=1}^{N_h} x_{hk} \qquad \text{(ratio estimator)}$$

where $\Sigma_{s_h} y_{hi}$, $\Sigma_{s_h} x_{hi}$ are sample totals of y and x in stratum h, and $\Sigma_{k=1}^{N_h} x_{hk}$ is the total of x in stratum h.

The population total is estimated by

$$\hat{T}^{*} = \Sigma_{h=1}^{H} \hat{T}_h$$

Royall and Herson [7] have shown that $\hat{T}^{*}$ is unbiased under $\xi(0,1: x)$ and is unbiased under more general polynomial models of degree J if the s_h are selected such that

$$\bar{x}_{s_h}^{(j)} = \bar{X}_h^{(j)}, \; j = 1, \ldots, J; \quad h = 1, \ldots H$$

Such a sample is called a stratified balanced sample of degree J and denoted by s*(J). Royall and Herson have proven that if $n_h \; \alpha \; N_h \bar{X}_h^{1/2}$ then under $\xi = \xi(\delta_0, \delta_1, \ldots, \delta_J: x)$ the strategy $[s^{*}(J), \hat{T}^{*}]$ is more efficient than $[s(J), \hat{T}_R]$.

If s is a sample which contains units from all strata, it can be shown that

$$\hat{T}^{*} = \Sigma_{h=1}^{H} \hat{T}_h$$

is optimal under a model ξ^{*} for which

$$Y_{hi} = \beta_h x_{hi} + \varepsilon_{hi}(x_{hi})^{1/2}$$

where $E_{\xi*}[\varepsilon_{hi}] = 0$ and $E_{\xi*}[\varepsilon^2_{hi}] = \sigma^2_h$, i.e. slope and variance scale factors vary from stratum to stratum. If a stratified balanced sample is chosen, then $\hat{T}^*$ is optimal under any model of the form

$$Y_{hi} = \delta_{0h}\beta_{0h} + \delta_{1h}\beta_{1h}x_{hi} + \dots + \delta_{Jh}\beta_{Jh}x^J_{hi} + \varepsilon_{hi}(x_{hi})^{1/2}$$

This says that $\hat{T}^*$, while designed to be optimal relative to model ξ^*, is robust relative to piecewise polynomial model departures.

While designed to generate approximately balanced samples, the basket method algorithm does a good job in generating approximate stratified balanced samples no matter how the strata boundaries are defined as long as the population strata sizes N_h are large relative to K. When a stratum size is small or moderate, it is difficult (if not impossible) to select a balanced sample of order J using any technique. For convenience, define strata sizes N_h so that $N_h/K = n_h$ are integers for h = 1, 2, ..., H. (It may happen that N_h/K may not be an integer but that is more of a nuisance than a problem in practice.) Then each basket sample s is an approximate stratified balanced sample of some order J since, by the nature of the basket algorithm,

$$\sum_{s_h} x^j_{hi} \doteq \frac{1}{K} \Sigma^{N_h}_{k=1} x^j_{hk}$$

j = 1, ..., J for some J and h = 1, ..., H. Thus

$$\bar{x}^{(j)}_{s_h} = \sum_{s_h} x^j_{hi}/n_h \doteq [(1/K)\ \Sigma^{N_h}_{k=1} x^j_{hk}/(N_h/K)] = \Sigma^{N_h}_{k=1} x^j_{hk}/N_h = \bar{X}^{(j)}_h$$

Also for s

$$\hat{T}^* = \Sigma_{h=1}^{H} [\Sigma_{s_h} y_{hi}/\Sigma_{s_h} x_{hi}) \Sigma_{k=1}^{N_h} x_{hk}]$$

$$\approx \Sigma_{h=1}^{H} [(\Sigma_{s_h} y_{hi}/(1/K) \Sigma_{k=1}^{N_h} X_{hk}) \Sigma_{k=1}^{N_h} X_{hk}]$$

$$= K \Sigma_{h=1}^{H} (\Sigma_{s_h} y_{hi}) = K \Sigma_{s} y_i$$

Since overall balance is achieved without regard to strata

$$\Sigma_{s} x_i \approx \frac{1}{K} \Sigma_{k=1}^{N} x_k$$

and thus

$$\hat{T}_R = K \Sigma_{s} y_i$$

Thus two estimators which are optimal under two different models become practically identical, for a basket produced sample, this estimator being optimal under a variety of models for this sample.

D. Relation to Classical Approach

1. P-unbiasedness of the basket strategy

We will refer to the pair $(P_{BSK}, \hat{T}_R)$ as the "basket strategy". This strategy is approximately P-unbiased since

$$E_{P_{BSK}} (\hat{T}_R) = \Sigma_{e=1}^{K} P_{BSK} (bsk(e)) \hat{T}_R$$

$$= \Sigma_{e=1}^{K} [(1/K)(\Sigma_{bsk(e)} y_i/ \Sigma_{bsk(e)} x_i) \Sigma_{k=1}^{N} x_k]$$

$$\approx \Sigma_{e=1}^{K} [(1/K)(\Sigma_{bsk(e)} y_i/(1/K) \Sigma_{k=1}^{N} x_k) \Sigma_{k=1}^{N} x_k]$$

$$= \Sigma_{e=1}^{K} (\Sigma_{bsk(e)} y_i) = \Sigma_{i=1}^{N} y_i = T$$

It could easily be made exactly p-unbiased if we changed P_{BSK} (bsk(e)) from 1/K to $\Sigma_{bsk(e)} x_i/\Sigma_{k=1}^{N} x_k$, that is, use selection probabilities proportional to basket totals. This would correspond to a PPAS plan over a very restricted class of samples.

We can also view the population as being made up of K "macro units" (baskets) from which a sample of size 1 is to be selected. Associated with macro unit k is an unknown number T_k, our goal being to estimate $T = \Sigma_{k=1}^{K} T_k$. Now the basket algorithm has the effect of making the component x-values in each basket resemble the x-values in the whole population, only "thinned out" somewhat, so that we should not expect the individual x-values in the selected basket to play a role in the estimation process. This is exactly what happens when we note that the ratio estimator (which uses the auxiliary information in the x-values) degenerates to the expansion estimator, that is

$$\hat{T}_R = (\Sigma_{bsk(f)} y_i / \Sigma_{bsk(f)} x_i)\, \Sigma_{k=1}^{N} x_k = KT_f = \hat{T}_E \quad \text{(when N = nK)}$$

where T_f is the y total in the randomly selected basket. The usefulness of the auxiliary information in the <u>process of estimation</u> has been eliminated during the <u>process of sampling</u>. But the sample <u>is random</u>, so that the objections of the proponents of the classical approach relative to purposive sampling do not apply to the basket strategy. The auxiliary information x, <u>assumed</u> to be relevant for inference, has been homogenized among baskets. Other variables assumed irrelevent, are dealt with through the randomization principle. Thus, the basket strategy should satisfy both the classicists and the advocates of the superpopulation approach.

2. Basket Sample as a Stratified Sample

Inherent in the basket procedure is a form of stratified sampling. Assume the population has been ordered in decreasing order and labelled 1, ..., N. Let S_1 denote the stratum of K units whose x-values are largest, S_2 the stratum of K units whose x-values are second largest, and so on until S_n where $n = [N/K]$. S_n contains the K + t smallest units where $t = N - nK$. In other words, all strata contain K units except possibly S_n which may contain K + t. If N/K is an integer, all are of size K. Every basket sample formed in the initial process of the algorithm contains one unit from each of $S_1, \ldots, S_n$ with possibly two units from S_n. The swapping routine may upset this structure slightly by trying to bring the totals into better balance. Regardless, each sample is a stratified sample, although not all stratified samples are possible due to the fixed number of bas:ets and the deterministic aspect of basket formation . Nevertheless, the samples are indeed representative in terms of the x-values.

3. Comparison with Simple Random Sampling and Stratified Random Sampling

For simplicity, assume that $N = nK$ and that the population is stratified as in section 2 so that each S_i contains K units. The variances of three sampling strategies will be compared

a. (SRS, $\hat{T}_E$): a simple random sample s of size n is drawn from the population (without regard to strata) and T is estimated by $\hat{T}_E = N\bar{Y}_s$.

b. (STRS, $\hat{T}_{ST}$): STRS indicates stratified random sampling. A stratified random sample s of size n is selected by choosing one unit at random from each of the n strata. The total T is estimated by

$$\hat{T}_{ST} = \Sigma_{h=1}^{n} K\bar{y}_h = K \Sigma_{h=1}^{n} y_h$$

where $\bar{y}_h$ is the sample mean in stratum h which equals y_h since only one unit is chosen from each stratum.

c. $(P_{BSK}, \hat{T}_R)$: the "basket method"

Let $V_P(\hat{T})$ denote the variance of the estimator $\hat{T}$ under the design P. Then for a specific population $y_1, \ldots, y_N$, it is well known that

$$V_{SRS}(\hat{T}_E) = (N(N-n)/n) \Sigma_{i=1}^{N} (y_i-\bar{T})^2/(N-1)$$

and

$$V_{STRS}(\hat{T}_{ST}) = K \Sigma_{j=1}^{n} \Sigma_{i=1}^{K} (y_{ij} - \bar{T}_j)^2,$$

where y_{ij} is the i observation in the jth strata and $\bar{T}_j$ is the jth stratum mean of y. Also,

$$V_{BSK}(\hat{T}_R) = \frac{N^2}{K} \Sigma_{e=1}^{K} (\bar{y}_{bsk(e)} - \bar{T})^2$$

since N = nk.

Direct comparison of $V_{SRS}(\hat{T}_E)$, $V_{STRS}(\hat{T}_{ST})$, and $V_{BSK}(\hat{T}_R)$ depends on the properties of the population at hand. There may be specific populations where each performs better than the others. Therefore, what we will do is to regard the population values $y_1, \ldots, y_n$ as being drawn from an infinite superpopulation described by a model. Hence the results obtained will not apply to any specific population but to the average of all populations that can be drawn from the superpopulation. Comparison will be made under two model formulations:

(i). Model I: $E_I[Y_i] = \mu$, $Var_I[Y_i] = \sigma_i^2$, $E_I(Y_i-\mu)(Y_j-\mu) = 0$ for any pair i, j.

Here we are hypothesizing that Y has no relationship with x and thus the ordering in the population is essentially random.

A modification of the results in Konijn [3] show that

$$\begin{aligned} E_I(V_{SRS}[\hat{T}_E]) &= E_I(V_{STRS}\ [\hat{T}_{ST}]) \\ &= E_I(V_{BSK}\ [\hat{T}_R]) \\ &= N(N-n) * \frac{1}{n} \Sigma_{i=1}^{N}\ \sigma_i^2/N \end{aligned}$$

Thus in this case the three strategies are equally efficient under the assumed model. As mentioned before for any one population there may be substantial differences between the three.

(ii) Model II: $E_{II}[Y_i] = \mu_i$, $Var_{II}[Y_i] = \sigma_i^2$, $E_{II}(Y_i-\mu_i)(Y_j-\mu_j) = 0$

Again slight modification of the results in Konijn [3] show that under this model

$$E_{II}[V_{SRS}\ (\hat{T}_E)] = N(N-n)(1/N)\ \Sigma_{i=1}^{N}\ \sigma_i^2/N + N(N-n)(1/n)\ \Sigma_{i=1}^{N}\ (\mu_i-\bar{\mu})^2/(N-1)$$

where $\bar{\mu} = \Sigma_{i=1}^{N}\ \mu_i/N$

$$E_{II}[V_{STRS}\ (\hat{T}_{ST})] = N(N-n)(1/n)\ \Sigma_{i=1}^{N}\ \sigma_i^2/N + (N^2/n)\ \Sigma_{j=1}^{n}\ \Sigma_{i=1}^{K}\ (\mu_{ij} - \bar{\mu}_{.j})^2/N$$

where μ_{ij} is the mean of the random variable Y_{ij} associated with the ith unit in stratum j and $\bar{\mu}_{.j} = \Sigma_{i=1}^{K}\ \mu_{ij}/K$, and

$$E_{II}[V_{BSK}\ (\hat{T}_R)] = N(N-n)(1/n)\ \Sigma_{i=1}^{N}\ \sigma_i^2/N + (N^2/K)\ \Sigma_{e=1}^{K}\ (\bar{\mu}_{bsk(e)} - \bar{\mu})^2$$

where $\bar{\mu}_{bsk(e)} = \Sigma_{bsk(e)}\ \mu_i/n$, the average of μ_i in basket e. Now let's suppose that $\mu_i = \alpha + \beta x_i$ and $\sigma_i^2 = \sigma^2 x_i$. Then substitution in the above formulas show that

$$E_{II}[V_{SRS}\ (\hat{T}_E)] \quad = A + N(N-n)(\beta^2/n)\ \Sigma_{i=1}^{N}\ (x_i-\bar{X})^2/(N-1)$$

$$E_{II}[V_{STRS}\ (\hat{T}_{ST})] \quad = A + (N/n)(K-1)\beta^2\ \Sigma_{j=1}^{n}\ \Sigma_{i=1}^{K}\ (x_{ij} - \bar{X}_{.j})^2/(K-1)$$

$$E_{II}[V_{BSK}\ (\hat{T}_R)] \quad = A + (N^2/K)\beta^2\ \Sigma_{e=1}^{K}\ (\bar{x}_{bsk(e)} - \bar{X})^2$$

$$\simeq A \qquad (\text{since } \bar{x}_{bsk(e)} \simeq \bar{X})$$

where

$$A = N(N-n)(\sigma^2/n)\bar{X}$$

Thus

$$E_{II}[V_{BSK}\ (\hat{T}_R)] \le E_{II}[V_{SRS}\ (\hat{T}_E)]$$

and

$$E_{II}[V_{BSK}\ (\hat{T}_R)] \le E_{II}[V_{STRS}\ (\hat{T}_{ST})]$$

The basket procedure is more efficient on the average from populations generated from Model II. Now $E_{II}[V_{SRS}\ (\hat{T}_E)]$ and $E_{II}[V_{STRS}\ (\hat{T}_{ST})]$ depend on the population variance of x and the variance within stratum of x. Thus determination as to which is more efficient depends on the particular x's at hand.

REFERENCES

1. Godambe, V. P. (1966). A new approach to sampling from finite populations I, II. J. R. Statist. Soc. B28, 310-328.

2. Godambe, V. P. and V. M. Joshi (1965). Admissibility and Bayes estimation in sampling finite population I. Ann. Math. Statist. 36, 1707-1722.

3. Konijn, H. S. (1973). Statistical Theory of Sample Survey Design and Analysis. London: North-Holland Publishing Co.

4. Royall, R. M. and Cumberland, William G. (1981). An empiral study of the ratio estimator and estimators of its variance. J. Amer. Statist. Ass. 76, 66-88.

5. Royall, R. M. and Eberhardt, Keith R. (1975). Variance estimates for the ratio estimator. Sankya, Ser. C 37, 43-52.

6. Royall, R. M.and Herson, J. (1973a). Robust estimation in finite population I. J. Amer. Statist. Ass. 68, 880-889.

7. Royall, R. M. and Herson, J. (1973b). Robust estimation in finite populations II: Stratification on a size variable. J. Amer. Statist. Ass. 68, 890-893.

8. Wallenius, K. T. and Belar, F. (1971). On Statistical Methods in Contract Negotiation, Part I. Technical Report NG, Department of Mathematics, Clemson University.

9. Wallenius, K. T. (1981). The Basket Methods for Selecting Balanced Samples - Part I: Theory. Technical Report No. N129, Department of Mathematical Sciences, Clemson University.

10. Wallenius, K. T. (1981). The Basket Method for Selecting Balanced Samples - Part II: Applications to Price Estimation. Technical Report No. 377, Department of Mathematical Sciences, Clemson University.

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE		READ INSTRUCTIONS BEFORE COMPLETING FORM
1. REPORT NUMBER N144	2. GOVT ACCESSION NO.	3. RECIPIENT'S CATALOG NUMBER
4. TITLE (and Subtitle) The Basket Method and its Role in Design and Model Oriented Approaches to Survey Sampling		5. TYPE OF REPORT & PERIOD COVERED Technical Report 6. PERFORMING ORG. REPORT NUMBER Technical Report #408
7. AUTHOR(s) R. J. Domangue and K. T. Wallenius		8. CONTRACT OR GRANT NUMBER(s) N00014-75-C-0451
9. PERFORMING ORGANIZATION NAME AND ADDRESS Clemson University Dept. of Mathematical Sciences Clemson, South Carolina 29631		10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS NR 047-611
11. CONTROLLING OFFICE NAME AND ADDRESS Office of Naval Research Code 411SP Arlington, Va. 22217		12. REPORT DATE January 1983 13. NUMBER OF PAGES 29
14. MONITORING AGENCY NAME & ADDRESS(if different from Controlling Office)		15. SECURITY CLASS. (of this report) Unclassified 15a. DECLASSIFICATION/DOWNGRADING SCHEDULE
16. DISTRIBUTION STATEMENT (of this Report) Approved for public release; distribution unlimited.		
17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)		
18. SUPPLEMENTARY NOTES This report superscedes report N129 "The Basket Method for Selecting Balanced Samples. Part I: Theory"		
19. KEY WORDS (Continue on reverse side if necessary and identify by block number) Balanced Sampling, ratio estimation, superpopulation approach		
20. ABSTRACT (Continue on reverse side if necessary and identify by block number) This paper discusses the basket method of sampling and estimation in finite populations in the framework of the traditional design based approach (randomization) and the model oriented approach (superpopulation). It is shown that the basket method incorporates the desirable features of both approaches. Its properties are studied.		

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

END

DTIC

8-86